Journal of Advances in Science and Technology
Vol. 18, Issue No. 2, September–2021, ISSN 2230-9659

A Review On Natural Language Processing's (Nlp's) Text Classification And Methods

Durga Gupta[1]*, Dr. Vijay Singh[2]

[1] Research Scholar, Shri Krishna University, Chhatarpur M.P.

[2] Associate Professor, Shri Krishna University, Chhatarpur M.P.

Abstract - Demands from business and society, as well as advances in technology, have been on the rise. The computer industry's technical progress is the end consequence of a lengthy chain of enormous and successful efforts by two primary forces: academics and industry. With NLP, machines may mimic human language comprehension. Natural language processing utilizes AI to take linguistic data, whether spoken or written, and transform it into a form that a computer can comprehend. Classifying raw texts into organized categories is called text classification. Email spam and legitimate correspondence may be separated using text categorization. This paper provides a clear breakdown of the many ways that natural language processing may be categorized. Based on current trends in technology acceptance and implementation, NLP is widely considered to represent the future of technology.

Keywords - Natural Language Processing, Language, Machine, Computer

----------X----------

1. INTRODUCTION

The term "computer" was first used in print in a book titled "The Young Man's Gleanings" in 1643. The abacus, also referred to as a "counting frame," is a tool for doing repetitive computations. Charles Babbage created the differential and analytical engine after a string of 19th-century breakthroughs. Inputting the data, doing the calculations, getting the answers on paper, and saving the results were all possible with this gadget. The structure is quite similar to that of a modern computer. That's why many see him as the pioneer of the computer age. Number crunching is a commonplace part of our daily lives, therefore we're all comfortable with it. When doing calculations, we rely on a wide variety of mathematical formulas and operations, such as addition, subtraction, multiplication, and division. There is a time savings associated with doing a simple computation. Time spent on sophisticated computations, however, is far more substantial. Similarly important is the precision of the computations. Therefore, man has ventured out with the notion to create a machine that can conduct all mathematical computation at a very high speed of operation with complete precision. This led to the development of the first computer. The verb "compute," meaning "to calculate," is the origin of the noun "computer."

2. LITERATURE REVIEW

Budiawan, Muhammad (2020) The following is a synopsis of the stated papers' contents, which may be categorized as theoretical computer science research. Given India's rich mathematical history, theoretical computer science is a vital area of study for computer scientists there. Several initiatives have been launched to increase the visibility of Indian scholars in the field of theoretical computing throughout the world. In addition, complexity theory is a major part of theoretical computer science. The number of steps in an algorithm is used to define it and determine how many different ways it can do a task. Over time, the complexity theory shifted its focus from the amount of time it took to calculate to the number of random bits required to encrypt a word securely. The employment of modulo prime and pseudo-prime numbers in cryptography has given rise to nontrivial notions that have been put to use in testing for primality. In contrast, Agrawal et al. of IIT Kanpur established in their work published in Annals of Mathematics that the derandomization problem may be addressed using breakthrough discoveries from PRIMES, which are in P. Algebraic complexity theory requires cooperation between computer science and algebra in the mathematical study of models like circuits. Algebraic techniques are used to demonstrate the difficulty of issues by showing lower limits. The next step is to understand that an ANN is a threshold-gated algebraic circuit. Weaker bounds were overcome thanks to increased knowledge of threshold gates. Theorists from India have also looked into NP-hard issues like isomorphism for structures. Tata Institute of Fundamental Research (TIFR) has made significant contributions to the study of communication complexity for problems involving the distribution of input to several entities. Additionally, Buchi was the

first to recognize the connection between logic and automata theory, and Pnueli later developed temporal logic as a language for describing reactive system properties. Model verification is hindered when reactive systems are seen as a mere sequential automaton. As a result, sequential interpretations of temporal logics are saddled with an exponential amount of interleaving for concurrent processes. The first temporal logic over traces, called TrPTL, was developed by CMI. It may be solved using a gossip automaton.

Syamil, Kamarul (2020) For the last three decades, India has been home to some of the most innovative minds in the field of theoretical computer science. One may also argue that throughout the 1980s and 1990s, when access to cutting-edge technology was limited, theory provided a great opportunity to keep up with international research in computers. A relatively recent area of study, parameterized algorithms and complexity analyze the execution of algorithms and develop new algorithms for challenging problems when combinatorial explosion is limited. India is closely linked to this thriving industry. In 1999, Chennai hosted the first ever international conference devoted only to this topic, and since then, India's pioneering Institute of Mathematical Sciences (IMSc) and Chennai Mathematical Institute (CMI) have been at the forefront of this field (CMI). Several significant efforts have been made by Indian scholars to characterize these issues in combinatorial terms, develop novel methods, and comprehend their parallel complexity. Indian scholars have contributed to network-level efforts to design data structures for static, concise representations and the maintenance of powerful information, as well as to the demonstration of non-minimalistic lower bounds on query complexity and space requirements. Circuit models, which include the computation of formal polynomials, often have considerable algebraic complexity. Due to the need to perform computations modulo prime and pseudo-prime numbers for a variety of cryptographic applications, error-correcting codes, and other fundamental computational issues, the practical importance of primality testing has skyrocketed in recent years, despite its long history as a topic of academic curiosity.

Kabanda, Gabriel (2019) This paper provides an analytical exposition, critical context, and integrative conclusion to the ongoing discussion of the meaning, significance, and potential applications of the theoretical foundations of computer science with regard to Algorithms Design and Analysis, Complexity Theory, Turing Machines, Finite Automata, Cryptography, and Machine Learning. Any clearly specified computing technique that accepts one or more values as input and generates one or more values as output is an algorithm. The finite control of a Turing machine is a finite program that can operate on a linear list of cells (the tape) with a single access pointer (the head). The collection of finite state machines known as a cellular automaton (inter-related). Assume that all computers are equivalent Turing machines. Any Turing machine's actions may be mimicked by the machine U. T. Automata were first offered as a basic model for the functioning of neurons because of their apparent similarity to this kind of computational structure. Computer scientists employ models of computation, which are mathematical abstractions of computers, to conduct rigorous research into computing. The terms "finite automaton" and "finite state machine" are used interchangeably to refer to the same thing: an automaton having a fixed number of states (FSM). The Church-Turing thesis says that any generic mathematical instrument for computing, such as a digital computer, is computationally comparable to the Turing machine. Efficiency, impossibility findings, approximation, the key role of randomness, and reductions are important topics in Theoretical Computer Science (TCS) (NP-completeness and other intractability results).

Dengel, Andreas (2018) This work explores the topic, "How may metaphorical representations in VR increase the knowledge of theoretical computer science concepts?" since metaphorical representations have been shown to have positive benefits in several areas of computer science. The notion of an interactive virtual classroom for teaching theoretical computer science concepts like finite state machines is introduced. A virtual reality prototype is being developed to transport pupils to a fantastical universe of islands (states) teeming with treasure islands (final states). Additional boats may be found on each island to represent each letter of the input alphabet. Students use a head-mounted display to gaze at a picture of one of the boats on the current island as they go from island to island. The player then presses a button on the controller to follow the chosen boat's shipping route (the state-transition function). It is up to the player to decide whether or not they want to finish the map or head straight for Treasure Island and the treasure box.

Oktar, Yigit& Turkan, Mehmet (2018) It has been suggested that a group of data clustering techniques, including an appropriate subspace search to locate inherent clusters, is the best course of action in the case of high dimensionality. There is a new spin on the subspace technique in sparsity-based clustering methods, which use an overcomplete dictionary representation to increase the number of dimensions. Therefore, these methods expand the scope of subspace clustering's applicability. On the other hand, clusters may be unstructured if the sparsity restriction is not included. Data clustering is achieved via the use of stronger restrictions, which may be thought of as a subset of grouping. An additional facet of sparsity-based clustering algorithms is the dictionary, the inverse of the sparsity constraint. Adaptive dictionaries, in contrast to fixed-waveform dictionaries, may be used to provide the state-model entity a more malleable shape. When coupled with organized sparsity, adaptive dictionaries coerce the state-model into coherent groups. Subspaces labeled with organized sparsity may be dissolved through

recursion to get deep sparse structures that map to a taxonomy. Finally, it's worth noting that this process may be further refined to include other machine learning angles.

3. NATURAL LANGUAGE PROCESSING AND ITS CLASSIFICATION

The capacity of a computer software to interpret spoken and written human language is known as natural language processing (NLP). Artificial intelligence includes this technique (AI). More than 50 years old, NLP has its origins in the study of language. Real-world uses range from scientific study to web indexing to BI. A branch of AI and linguistics, "natural language processing" (NLP) is focused on teaching computers to decipher human-written language. The goal of developing NLP was to allow users to more easily interact with computers by using more natural language. Since not every user will be fluent in machine-specific language, NLP is designed to help people who do not have the time to become experts in it. You may think of a language as a system of rules or a system of symbols. Information is sent or communicated via a combination of symbols. The Rules are like a tyrant over symbols. In its most basic form, Natural Language Processing (NLP) may be broken down into two subfields: the first, Natural Language Understanding, focuses on the analysis of text, while the second, Natural Language Generation, focuses on the creation of new text (Figure 1).

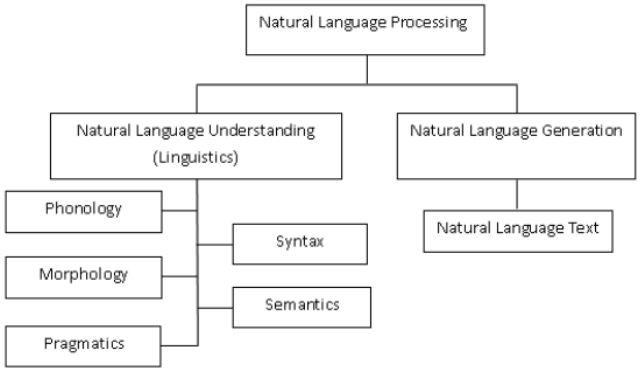

Figure 1: Classification of NLP

NLP Methods

Natural Language Processing (NLP) integrates AI, linguistics, and computer science. A large number of researchers worked on it, creating programs, instruments, and systems to deal with pressing issues and emerging difficulties. This section satisfies this need by discussing the most popular NLP methods used in academic studies, including SS, MT, TS, SA, NER, and ED. The purpose of this section is to lay out the many methods that may be used, the most popular packages that implement these methods, and the pros and cons of each.

- **Semantic Search:** Sentiment analysis (also known as opinion mining) is a kind of semantic search that use natural language processing (NLP) to ascertain if information is positive, negative, or neutral in tone. Performing sentiment analysis on textual data is a common practice for firms looking to track consumer feelings about a product or service, as well as better understand what those customers want.

- **Semantic Analysis:** Semantic analysis is the process of determining the intended meaning of a given sentence. The emphasis is on the literal sense of all words and phrases. Sentence structure is also addressed. The literal meaning of the text is taken from the dictionary. The text's meaning is dissected for analysis. This is done by creating a syntactic model of the items and structures found in the task domain.

- **Text Summarization:** Through the use of natural language processing, text summarizing simplifies complex information by reducing scientific, medical, technical, or other jargon to its essential concepts.

- **Keywords Extraction:** Natural language processing (NLP) is a method for analyzing texts that includes keyword extraction, often known as keyword identification or keyword analysis. With this method, you may have it automatically pull out the most used words and phrases from the main body of a document. It is often the initial stage in summarizing a text and communicating the core ideas and messages of the source material.

- **Text Embedding:** Word embedding, a kind of text embedding, is a method of representing words wherein words with similar meanings are given comparable representations. They are a text representation in a distributed format, and they may be a substantial advance for deep learning techniques' impressive performance on challenging natural language processing tasks. Through text corpus word embedding techniques, a real-valued vector representation of a fixed-size vocabulary may be learned.

- **Named Entity Recognition:** Because we in the IT industry love our acronyms so much, we often refer to the Natural Language Processing approach of identifying and extracting "named identities" from text as "Named Entity Recognition" (or "NER").

- **Emotion Detection:** A person's core behaviors may be analyzed with the help of their feelings. A wide variety of sources, such as words said or written, might

contribute to this data. It accurately detects sarcasm and other forms of subjective emotion in written text.

Natural Language Processing (NLP) applications

- **Email filters:** One of the first and most fundamental uses of NLP in cyberspace is email filtering. It all began with spam filters, which look for certain patterns in messages that indicate they are spam. But just as early iterations of NLP were improved upon, so too have filtering practices. Gmail's inbox organization is one of the most recent and widespread uses of natural language processing. Based on their contents, the system classifies emails as main, social, or promotional. All Gmail users may benefit from this feature since it helps them focus on the most pressing messages and reduces the amount of time spent sifting through irrelevant messages.

- **Search results:** To help the typical user locate what they need without having to be a search-term wizard, search engines utilize natural language processing to surface relevant results based on comparable search habits or user intent. By looking at the whole picture and understanding what you mean rather than the precise search phrases, Google is able to not only guess what popular searches may apply to your query as you start typing but also provide more relevant results.

- **Predictive text:** Features like autocorrect, autocomplete, and predictive text have become so ubiquitous on modern smartphones that we hardly think twice about using them. Comparable to search engines, autocomplete and predictive text fill incomplete words or propose related alternatives as you enter. Sometimes autocorrect may even modify individual words to improve the flow of the sentence.

- **Language translation:** If your Spanish assignment is grammatically incorrect, it's a good indication that you didn't do it on your own. It's been common practice for translation services to gloss over the fact that many languages have various sentence structure ordering that prevent literal translation. And they have progressed considerably. Using natural language processing (NLP), online translators may provide more precise and grammatically sound translations.

- **Digital phone calls:** Phone calls in digital form: While we've all heard the phrase "this call may be recorded for training reasons," few of us have ever stopped to consider what that means. It turns out that these recordings are usually stored in a database for an NLP system to learn from and better in the future, however they may be utilized for training reasons if a client is upset.

- **Data analysis:** More business intelligence (BI) providers now provide a natural language interface to data visualizations, allowing analysts to conduct analyses directly in their preferred language. Better visual encodings are one such example; they provide the most appropriate visualization for a given job by factoring in the data's semantics.

4. TEXT CLASSIFICATION BY USING NATURAL LANGUAGE PROCESSING (NLP)

Classification of text is a technique that encompasses other names, such as text tagging and text categorization. Text classifiers use Natural Language Processing (NLP) to automatically evaluate text and classify it into a series of pre-defined tags or categories. There have been significant advancements in AI without any required changes to the underlying technology. An AI application may be run on a vintage PC. Conversely, machine learning's potential for good is almost boundless. Natural Language Processing is a subfield of artificial intelligence that enables robots to read, comprehend, and communicate ideas. Healthcare, the media, finance, and human resources are just few of the fields where NLP has been a huge success. Texts and speeches are the most typical examples of unstructured data. There is a lot of it, but it's not easy to get to the good stuff. Absent this, data mining would be a lengthy process. Information abounds in both written and spoken forms. It's because we rely on written and spoken language as our major means of interaction with the world.

Sentiment analysis, cognitive assistance, span filtering, the detection of bogus news, and instantaneous language translation are just some of the tasks that may be performed by NLP on this data. Because of the growth of the Internet, information can now travel much more quickly throughout the globe. Although technology broadens access to knowledge, it also makes individuals more susceptible to spreading fabricated stories and gossip. Fake news about a company can have a significant impact on its stock price, and fake news posted during a disaster, such as Hurricane Irma in 2016 or the current Covid-19 pandemic, can cause unnecessary tension among people and, in the worst case, put lives at risk by preventing people from correctly recognizing the severity of the situation. Therefore, it has been an essential social purpose to taint the online dissemination of bogus news.

Approaches

There are three primary methods for classifying texts:

i. **Rule-based approaches:** Those methods that rely on rules to categorize text are called rule-based. One technique for classifying texts is to build a list of keywords associated with a certain column and then evaluate texts based on the frequency with which those keywords appear. Words like "fur," "feathers," "claws," and "scales," for instance, can assist a zoologist locate internet materials pertaining to animals. These methods are tedious to implement, may take a long time, and are hard to scale.

ii. **Machine learning approaches:** Machine learning allows us to train models on enormous collections of text data, which then allows us to make predictions about the categories that fresh text will fall into. Feature extraction is the process of converting textual information into numerical information for use in training models. Bag of words and n-grams are two important feature extraction methods. It is possible to utilize a number of effective machine learning techniques for text categorization. Here are a few of the most common:

   ✓ Naive Bayes classifiers

   ✓ Support vector machines

   ✓ Deep learning algorithms

iii. **Hybrid approaches:**The aforementioned two algorithms have been merged into one in these methods. They create a classifier that can be modified for specific situations using rule-based and machine learning approaches.

Applications of text classification

Text categorization has many further uses across many fields than those already listed.

- Classifying texts is useful for determining a text's language, such as when trying to figure out what people are tweeting or posting in. Google Translate, for instance, can tell what language you're speaking automatically.

- Forensic analysis and literary studies are only two of the many areas that put text categorization to use to determine the true authors of previously unattributed writings.

- Recently, text categorization has been utilized for triaging messages in a mental health services online support forum. Every year, people in the natural language processing (NLP) field hold contests (see: clpsych.org) to find solutions to text categorization difficulties arising from clinical research.

5. CONCLUSION

We can say that classification problems may be broken down into several subproblems, one of which is text classification, in which the input data point(s) is text and the aim is to place the text into one or more buckets (called classes) from a collection of pre-defined classes (classes). ICT, or information and communication technology, is now used in every industry and area of study. In this article, we'll go through the fundamentals of computing, including the various computer architectures. In addition, the idea of algorithms is discussed in great detail as they pertain to MU. Also included is a generational breakdown of the computing industry. Natural Language Processing (NLP) is a subfield of AI that builds models for data extraction and text comprehension using a variety of ML techniques, proprietary NLP methods, and DL. In addition to improving efficiency and accuracy on the work, NLP also simplifies people's daily life. Recent years have seen a meteoric rise in its popularity among academics.

REFERENCES

1. Budiawan, Muhammad. (2020). Report: Research on Computer Science Theory.

2. Syamil, Kamarul. (2020). Research In Theoretical Computer. 9.

3. Kabanda, Gabriel. (2019). On the Theoretical Foundations of Computer Science. An Introductory Essay.

4. Dengel, Andreas. (2018). Seeking the Treasures of Theoretical Computer Science Education: Towards Educational Virtual Reality for the Visualization of Finite State Machines. 10.1109/TALE.2018.8615288.

5. Oktar, Yigit& Turkan, Mehmet. (2018). A Review of Sparsity-based Clustering Methods. Signal Processing. 148. 10.1016/j.sigpro.2018.02.010.

6. Carl Hamacher, Computer Organization, Fifth Edition, Asia, 2013.

7. P. K. Sinha, P.Sinha, Fundamentals of Computers, BPB Publishers, 2007.

8. Chhajed. N, U. Imran, Simarjeet. S, B., “A comparision Based analysis of Four Different Types of sorting Algorithms in data structures with Their Performances, International Journal of Advanced Research

in Computer Science and software Engineering, vol.3, issue.2, February 2013.

9. A. Jahad, M. Rami, "An enhancement of major sorting algorithms", The International Arab Journal of Information Technology, vol.7, no.1, 2010.

10. Tapaswi, N., & Jain, S. (2012, September). Treebank based deep grammar acquisition and Part-Of-Speech Tagging for Sanskrit sentences. In Software Engineering (CONSEG), 2012 CSI Sixth International Conference on (pp. 1-4). IEEE.

Corresponding Author

Durga Gupta*

Research Scholar, Shri Krishna University, Chhatarpur M.P.